

调研报告

——基于文本挖掘方法的智能推荐等功能在农业领域中的实现

一、 背景概述

文本挖掘是一个新的数据挖掘领域，试图从自然语言中提取出有价值的信息。然而，区别于一般的数据挖掘，文本挖掘需要从有限结构化甚至是非结构化的文本数据抽取散布的知识。我们希望，能够通过快速、有效的文本挖掘方法，实现语义模式的搜索，而不仅仅是搜索出关键字。

而我们的案例所面临的问题就是：如何将一些已有的文本挖掘方面的理论基础及实际方法，应用到农业领域，即在我们已有的有关“病虫害”、“生产技术”等农业科普知识文档中，实现自动化处理文档、智能推荐等功能。

二、 理论解决方案

文本挖掘的基本过程如下图所示，主要分为：文本预处理、特征提取、文本挖掘、评估四个阶段。

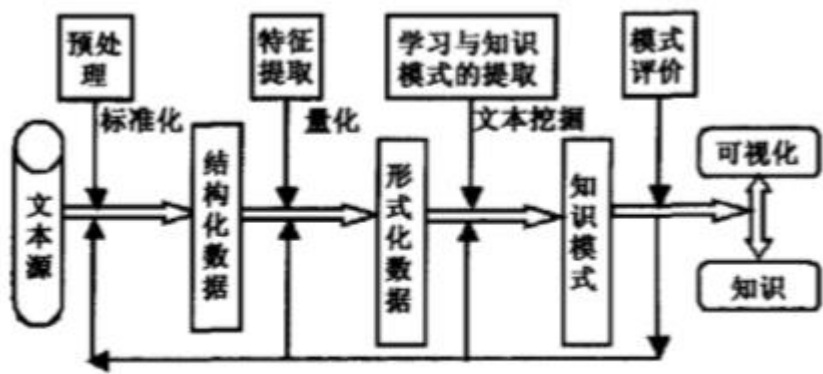


图 1 文本挖掘的过程示意图

（一） 文本预处理

文本预处理又可以分为 3 个步骤：

1. 文本清理：将文本中多余的广告等信息过滤去除。在我们的案例中，农业科普知识文档均为现成，所以不需要进行这一步骤。
2. 文本分块：将每篇文档根据空白处与某些标点等特征分为几块。
3. 文本标注：对上述已进行分块的文档中的每一块内容用关键词进行标注。标注内容需要从相关词库中检索，并根据上下文相关规则选取合适的标签进行标注。

（二） 特征提取

在文本数据存在多余（不能提供额外信息的特征）或是不相关（不能提供有用信息的特征）特征的前提之下，将文本的重要、有用的特征提取出来，建立成为模型。常用的模型有：布尔逻辑模型，向量空间模型(VSM)，潜在语义索引(LSI)和概率模型；其中应用较多且效果较好的是向量空间模型。

（三） 文本挖掘

目前研究和应用最多的几种文本挖掘技术有：文档聚类、文档分类和自动文摘。

1. 文档聚类：

首先，文档聚类可以发现与某文档相似的一批文档，帮助知识工作者发现相关知识；其次，文档聚类可以将一个文档聚类成若干个类，提供一种组织文档集合的方法；再次，文档聚类还可以生成分类器以对文档进行分类。

文本挖掘中的聚类可用于：提供大规模文档集内容的总括；识别隐藏的文档间的相似度；减轻浏览相关、相似信息的过程。

聚类方法大致可以分为：层次聚类法与平面划分法。

前者是最常用的群集方法，它能够生成层次化的嵌套簇，且准确度高。但是在每次合并时，需要全局地比较所有簇之间的相似度，并选择出最佳的两个簇，因此执行速度太慢，不适合大量文本聚类。

后者将文本集合水平地分割为若干个簇，而不是生成层次化的嵌套，因此，该方法的执行速度较快，但是必须事先确定 k 的取值，且种子选取的好坏对聚类结果有较大影响。

常用的聚类划分方法有 K -平均算法和 K -中心算法。

2. 文档分类：

分类和聚类的区别在于：分类是基于已有的分类体系表的，而聚类则没有分类表，只是基于文档之间的相似度。

文档分类可用于：用户刚开始接触一个领域想了解其中的情况；用户不能够准确地表达自己的信息需求时；减少用户分析检索结果的工作量

文档自动分类一般采用统计方法或机器学习来实现。常用的方法有：简单贝叶斯分类法， K -最近邻参照分类算法以及支持向量机分类方法等。

3. 自动文摘：

从文章中自动抽取关键句或关键段，即人类的理解是能够概括文章中心的句子或段落。机器的理解只能模拟人类的理解，即拟定一个权重的评分标准，给每个句子打分，之后给出排名靠前的几个句子。具体方法请见：

<http://www.hankcs.com/nlp/textrank-algorithm-java-implementation-of-automatic-abstract.html>

（四）评估

评估之后，结果可以被丢弃或者作为下一个序列集的输入。评估的标准有：

1.查全率和查准率：前者为实际被检出的文本的百分比，后者为检索到的实际文本与查询相关文本的百分比。两者很难同时很好，需要权衡。

2.冗余度和放射性：前者为信息抽取中冗余的程度，后者为抽取增多时错误增多的趋势。一般保留一定冗余度来降低放射性。

三、 类似案例

An example of biomedical text mining

Two parts:

- MedLEE (Medical Language Extraction and Encoding)

- to extract, structure and encode clinical information so that it can be further exploited by subsequent automated processes

- cTAKES (clinical Text Analysis and Knowledge Extraction System)

- consists of the following NLP modules: sentence boundary detector, tokenizer, normalizer, POS tagger, shallow parser and NER annotator (including status and negation annotators)

- examples:

- UIMA CVD: Process raw text and view NLP metadata;

- UIMA CPE: Process a multiple document batch;

- cTAKES GUI: Process raw text in a web browser;
- TimeLane: View extracted UMLS and Temporal information
- <http://ctakes.apache.org/>
- URL: <http://www.sciencedirect.com/science/article/pii/S1386505614001105>

四、 相关工具

（一些可能用到的开源包）

1、HanLP（汉语言处理包）

- 由一系列模型与算法组成的 Java 工具包，完全开源
- 有中文分词、命名实体识别、摘要关键字、智能推荐、自动摘要、句法分析和简繁体转换等功能。
- 相关用法和算法的详细介绍以及 Q&A 见下网址：
<http://www.hankcs.com/nlp/hanlp.html>
- 官网: <http://hanlp.linrunsoft.com>

2、FudanNLP

- 为中文自然语言处理而开发的工具包
- 可以实现：中文处理（分词、词性标注等）、信息检索（文本分类、新闻聚类）和机器学习 (Average Perceptron、K-means、Exact Inference、Passive-aggressive Algorithm)
- 官网: <http://www.oschina.net/action/project/go?id=14707&p=home>

3、OpenNLP

- 是一个基于 Java 机器学习工具包
- 支持大多数常用的 NLP 任务，如：标识化、句子切分、部分词性标注、名称抽取、组块、解析等
- 官网: <http://opennlp.apache.org/>

五、 总结

1. 在网上搜索较完备的农业专业词库（可以结合输入法词库），和基础词库合并构成所需词库，并协调两者间的冲突。
2. 因为有一些名称因地区差异而不同，需要完善同义词库和相关词库。
3. 将词库导入，利用一些现有库编码，以达到自动化处理文档和智能推荐等功能。
4. 之后需要对程序进行多次测试，以排除自带的和合成不兼容问题，对一些库来说还可以促进机器学习，从而提高精确度。